LLM-QAT: Data-Free Quantization Aware Training for Large Language Models

Year: 2023 | Citations: 281 | Authors: Zechun Liu, Barlas O■uz, Changsheng Zhao

---

Abstract

Several post-training quantization methods have been applied to large language models (LLMs), and have been shown to perform well down to 8-bits. We find that these methods break down at lower bit precision, and investigate quantization aware training for LLMs (LLM-QAT) to push quantization levels even further. We propose a data-free distillation method that leverages generations produced by the pre-trained model, which better preserves the original output distribution and allows quantizing any generative model independent of its training data, similar to post-training quantization methods. In addition to quantizing weights and activations, we also quantize the KV cache, which is critical for increasing throughput and support long sequence dependencies at current model sizes. We experiment with LLaMA models of sizes 7B, 13B, and 30B, at quantization levels down to 4-bits. We observe large improvements over training-free methods, especially in the low-bit settings.